

## Extracting Text from Scanned PDFs

If you are using Adobe Acrobat Standard or Pro, you can extract and work with editable text from a *scanned* PDF.

Remember, PDF files can be made

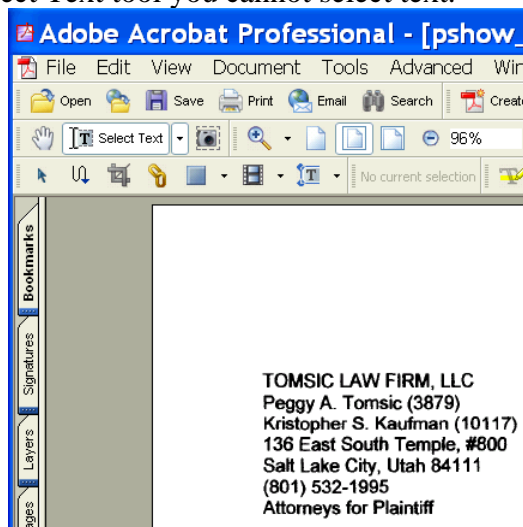
- by selecting a “PDF creating” printer after you create your word processing document, or by using the WordPerfect “Publish to PDF” command or the PDF button in Microsoft Word. This method of creating a PDF results in a file that you can search for text, and from which you may copy and paste text. You may also SAVE AS, selecting Microsoft Word format or RTF format, and extract most of the text from the file (though formatting may be a little funky. Because of these features, this sort of PDF document is often called a “text” PDF.

OR

- by scanning a document, with a scanner or photocopier with “scan to PDF” ability. But if a PDF file is created by this method – scanning – then the file cannot be searched, copied from or SAVED AS to Word or RTF format. This sort of PDF document is often called an “image” PDF.

This document explains explains how to get text finding, copying and outputting features in a *scanned* PDF document.

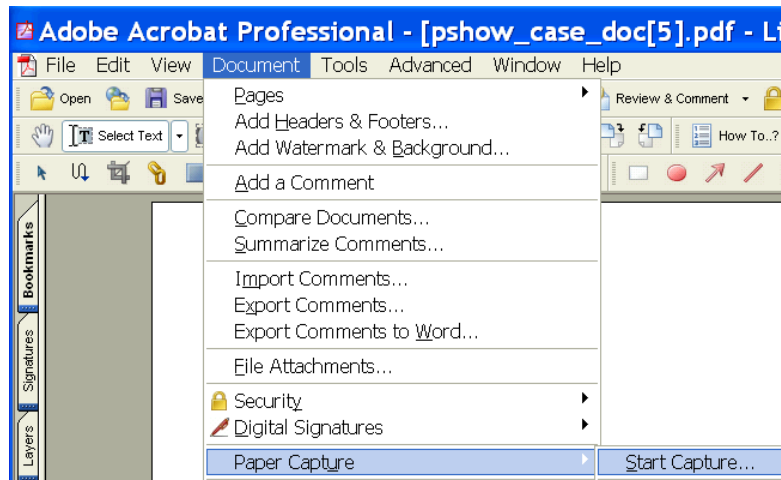
Open the document. Make sure it is a “scanned PDF” by attempting to select text. Even with the Select Text tool you cannot select text.



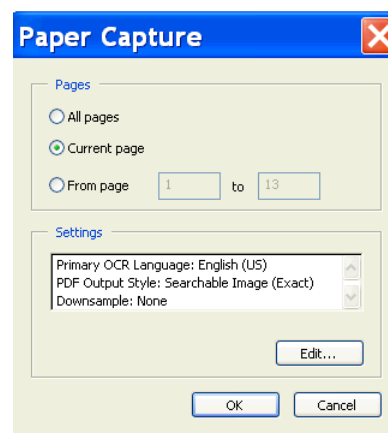
[If you are working with a document in CM/ECF, be sure you are working *without headers*. Headers will prevent this method from working. In [QuickQuery](#), headers do not show.]



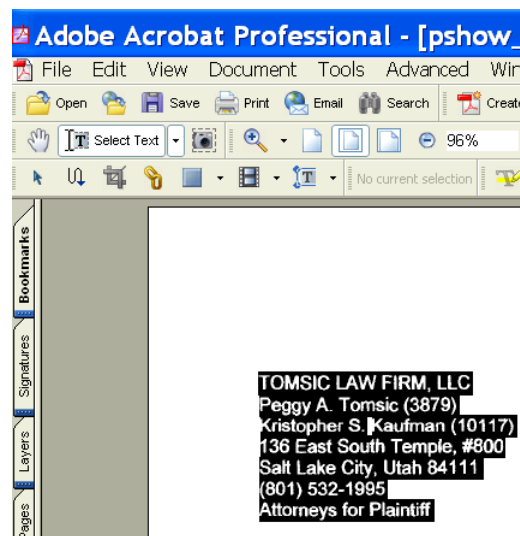
Then, with the document open, on the Document Menu, pick Paper Capture, Start Capture:



In the dialog that appears select as much of the document as you want to work with in text. In this case I only wanted the current page:



Now, the text selection tool will work with the document, for as much of it as you “captured.” If you capture the entire document, you can search it all. You may also SAVE AS an RTF file or Microsoft Word document.



My experience has shown this feature results in very accurate conversions, though there will be line breaks or “hard returns” that need to be removed.